INFORMATION CONTENT IN MEDLINE RECORD FIELDS

Ronald N. Kostoff, Ph. D.
Office of Naval Research
Arlington, VA 22217

Joel A. Block, M.D.
Rush Medical College
Chicago, IL 60612

Jesse A. Stump
Office of Naval Research
Arlington, VA 22217

Kirstin M. Pfeil
Office of Naval Research
Arlington, VA 22217

(The views in this paper are solely those of the authors, and do not represent the views of the Office of Naval Research, or Rush Medical College)

ABSTRACT

Background: The authors have been conducting text mining analyses (extraction of useful information from text) of Medline records, using Abstracts as the main data source. For literature-based discovery, and other text mining applications as well, all records in a discipline need to be evaluated for determining prior art. Many Medline records do not contain Abstracts, but typically contain Titles and Mesh terms. Substitution of these fields for Abstracts in the non-Abstract records would restore the missing literature to some degree.

Objectives: Determine how well the information content of Title and Mesh fields approximates that of Abstracts in Medline records.

Approach: Select historical Medline records related to Raynaud's Phenomenon that contain Abstracts. Determine the information content in the Abstract fields through text mining. Then, determine

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **01 JUN 2004**	2. REPORT TYPE	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **INFORMATION CONTENT IN MEDLINE RECORD FIELDS**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **RONALD KOSTOFF; JOEL BLOCK; JESSE STUMP; KIRSTIN PFEIL**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **OFFICE OF NAVAL RESEARCH,800 N. QUINCY STREET,ARLINGTON,VA,22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Background: The authors have been conducting text mining analyses (extraction of useful information from text) of Medline records, using Abstracts as the main data source. For literature-based discovery, and other text mining applications as well, all records in a discipline need to be evaluated for determining prior art. Many Medline records do not contain Abstracts, but typically contain Titles and Mesh terms. Substitution of these fields for Abstracts in the non-Abstract records would restore the missing literature to some degree. Objectives: Determine how well the information content of Title and Mesh fields approximates that of Abstracts in Medline records. Approach: Select historical Medline records related to RaynaudŸs Phenomenon that contain Abstracts. Determine the information content in the Abstract fields through text mining. Then, determine the information content in the Title fields, the Mesh fields, and the combined Title-Mesh fields, and compare with the information content in the Abstracts. Results: Four metrics were used to compare the information content related to RaynaudŸs Phenomenon in the different fields: total number of phrases; number of unique phrases; content of factors from factor analyses; content of clusters from multi-link clustering. The Abstract field contains almost an order of magnitude more phrases than the other fields, and slightly more than an order of magnitude more unique phrases than the other fields. Each field used a factor matrix with fourteen factors, and the combination of all 56 factors for the four fields represented 27 separate, but not unique, themes. These themes could be placed in two major categories, with two sub-categories per major category: Auto-immunity (antibodies, inflammation) and circulation (peripheral vessel circulation, coronary vessel circulation). All four sub-categories included representation from each field. Thus, while the focus of the representation of each field in each sub-category was moderately different, the four sub-category structure could be identified by analyzing the total factors in each field. In the cluster comparison phase of the study, the phrases used to create the clusters were the most important phrases identified for each factor. Thus, the factor matrix served as a filter for words used for clustering. While clusters were generated for all four fields, the Title hierarchy tended to be fragmented due to sparsity of the co-occurrence matrix that underlies the clusters. Therefore, the Title clusters were examined at only the lower levels of aggregation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**40**	

the information content in the Title fields, the Mesh fields, and the combined Title-Mesh fields, and compare with the information content in the Abstracts.

Results: Four metrics were used to compare the information content related to Raynaud's Phenomenon in the different fields: total number of phrases; number of unique phrases; content of factors from factor analyses; content of clusters from multi-link clustering. The Abstract field contains almost an order of magnitude more phrases than the other fields, and slightly more than an order of magnitude more unique phrases than the other fields.

Each field used a factor matrix with fourteen factors, and the combination of all 56 factors for the four fields represented 27 separate, but not unique, themes. These themes could be placed in two major categories, with two sub-categories per major category: Auto-immunity (antibodies, inflammation) and circulation (peripheral vessel circulation, coronary vessel circulation). All four sub-categories included representation from each field. Thus, while the focus of the representation of each field in each sub-category was moderately different, the four sub-category structure could be identified by analyzing the total factors in each field.

In the cluster comparison phase of the study, the phrases used to create the clusters were the most important phrases identified for each factor. Thus, the factor matrix served as a filter for words used for clustering. While clusters were generated for all four fields, the Title hierarchy tended to be fragmented due to sparsity of the co-occurrence matrix that underlies the clusters. Therefore, the Title clusters were examined at only the lower levels of aggregation.

The Abstract, Mesh, and Mesh + Title fields had the same first level taxonomy categories, autoimmunity and circulation. At the second level, the Abstract, Mesh, and Mesh + Title fields had the autoimmune diseases and antibodies sub-category in common. The Abstract and Mesh fields shared fascia inflammation as the other autoimmunity sub-category, while the other Mesh + Title sub-category focuses on vinyl chloride poisoning from industrial contact, and consequences of antineoplastic agents. However, in both cases, even though the words may be different, inflammation may be the common theme.

Conclusions: For taxonomy generation, especially at the higher levels, each of the four fields has a similar thematic structure. At very detailed levels, the Mesh and Title fields run out of phrases relative to the Abstract field. Therefore, selection of field (s) to be employed for taxonomy generation depends on the objectives of the study, particularly the level of categorization required for the taxonomy.

For information retrieval, or literature-based discovery, selection of the appropriate field again depends on the study objectives. If large queries, or large numbers of concepts or themes are desired, then the field with the largest number of technical phrases would be desirable. If queries or concepts represented by the more accepted popular terminology is adequate, then the smaller fields may be sufficient. Because of its established and controlled vocabulary, the Mesh field lags the Title or Abstract fields in currency. Thus, the Title or Abstract fields would retrieve records with the most explicitly stated current concepts, but the Mesh field would capture a larger swath of fields that contained a concept of interest but perhaps had a wider range of specific terminology in the Abstract or Title text.

In addition, this study provides the first validated estimate of the disparity in information retrieved through text mining limited to Titles and Mesh terms relative to entire Abstracts. As much of the older biomedical literature was entered into electronic databases without associated Abstracts, literature-based discovery exercises that search the older medical literature may miss a substantial proportion of relevant information. On the basis of this study, it may be estimated that up to a log order more information may be retrieved when complete Abstracts are searched.

BACKGROUND

Text mining is the extraction of useful information from text (1-3). In its modern day incarnation, it refers to information technology-based extraction from very large volumes of text. One variant of text mining is literature-based discovery (4-8). It generates actual discovery from text only, using intermediate literatures that link the problem literatures (literatures that describe the problem to be solved) to discovery literatures (literatures that contain possible solutions to the problem).

The authors have been conducting a literature-based discovery study of Raynaud's Phenomenon. The first phase of this study is to compare the authors' results with those of Swanson's pioneering literature-based discovery paper on Raynaud's Phenomenon (6), and with subsequent papers that attempted to replicate Swanson's results (7-8).

Medline has been used as the source database for all of the above Raynaud's Phenomenon studies. The present study also used Medline Abstracts as the data source. However, in the 1975-1985 time frame that immediately preceded Swanson's initial Raynaud's Phenomenon publication, only 58% of the Medline records relating to Raynaud's Phenomenon contained Abstracts. Since a critical element of discovery requires establishing that prior art does not exist, all prior records need to be evaluated. Some type of information from these non-Abstract records needs to be included in the empty Abstract field. The only other type of text-based information available from Medline is contained in the Title and Mesh fields. This leads to the question: how well does the information content in the Title and Mesh fields approximate that in the Abstract fields?

Most of the studies comparing text fields in a full text document or database record have the objective of comparing information retrieval. In examining different approaches to indexing medical literature, Hersh and Hickam (9) found that text word indexing is more effective than MESH term indexing. In a test of whether Abstract words occurred as frequently as could be expected from full text statistical analysis, Su et al (10) found that 96-98% of the Abstracts tested were not significantly different from random samples of the articles they represented. However, from a retrieval perspective, Johnson et al (11) found that searching the full text version of NEJM retrieved a larger number of records than Medline. This was due primarily to methodology terms found in the text but not in the Title or Abstract. Additionally, in Medline, two indexer-supplied fields (descriptors and publication types) retrieved 11-89% more than Title or Abstract alone. This important role played by MESH in enhancing information retrieval was supported by Srinivasan (12). However, while MESH terms can enhance information retrieval, their limitations include the Indexer Effect (13), indexer consistency in Medline (14), indexer-searcher vocabulary mis-matches (15), and lag in adopting new terminology. Even concept hierarchies to aid the management of controlled vocabularies, such as the Metathesaurus of the Unified Medical Language System (16-17), may not always incorporate evolving topics in a timely manner (18).

Mijnhout et al (19) concluded that, for comprehensive retrieval, both MESH terms and text words should be used in a search strategy, implying a disparity between the fields. He also concluded that, for comprehensiveness, multiple databases should be searched. None of these articles have focused on information content exclusively, and used the combined clustering and unique phrase occurrence approach that will be described in this paper.

OBJECTIVES

This study will determine how well the information content in the Title and Mesh fields of Medline records approximates that contained in the Abstract field.

APPROACH

General

The approach consists of two components: quantitative and qualitative. The quantitative component compares total and unique words in each field. The qualitative component compares category taxonomies in each field. The former provides detail, while the latter provides structure.

Specific

The query “RAYNAUD’S DISEASE OR RAYNAUD*[TW]” (restricted to 1975-1985) was inserted into the PubMed Search engine for Medline, and retrieved 932 records with Abstracts (these are the 58% of the total records in the 1975-1985 time frame that contained

Abstracts). The contents of the Abstracts field, Titles field, Mesh field, and Mesh-Titles field were placed in separate databases. The information content in each of these four databases (Abstract, Title, Mesh, and combined Title-Mesh) was compared using text mining. Both qualitative and quantitative metrics were used for the comparison.

As with any quantitative comparison procedure, a critical item is the selection of appropriate metrics. In comparing any two databases for this study, the main focus is on assessing the importance of unique phrases and phrase patterns in each database relative to the other databases. This assessment is made at two levels of aggregation.

At the lower phrase-focused level, the numbers of total and unique phrases, and the significance of each phrase, are compared among the databases. Because of the present study's context of information for literature-based discovery, and literature-based discovery's extensive use of word/ phrase matching among documents from different retrievals, numbers of phrases becomes significant in the matching process. The greater the number of phrases, the larger the dimensions describing each concept, and the greater the probability that two equivalent concepts will exhibit some overlap of their component phrases. At the higher phrase-pattern focused level, the phrase clusters and overall taxonomies are compared among the databases.

1) Lower Level Comparison

The <u>number of phrases</u> is relatively straight-forward, although not quite as simple as it would appear superficially. It is easy to count phrases produced by a Natural Language Processor, or other type of phrase generator. It is more complex to count the subset of generated phrases that have high technical content, and that would be used in the performance of an actual text mining study. Judgement is required to translate the raw phrase generator outputs to useful phrases.

Even the metrics for classifying a phrase as "high technical content" are ill-defined. High technical content is a function of context, and classifying a phrase in isolation from its context is fraught with error.

In the phrase clustering process presented later in this paper, a method called factor matrix filtering is used to insure that high technical content phrases only are used for clustering. Only those phrases that have substantial influence on determining the themes of the factors in a factor analysis are used for the clustering, and these phrases become high technical content by virtue of their function and context.

The significance of each phrase is far more complex to obtain, since it depends on the context of the analysis to be performed. If a macro level analysis is the objective, such as development of a higher level taxonomy, then a database that is deficient in a few phrases relative to some other more detailed database may affect the final taxonomy relatively little (hypothetical at this point). However, if a micro level analysis is the objective, such as literature-based discovery (4-8), then a deficiency of a very few phrases may be crucial to the results, if the discovery elements are contained within these missing phrases.

The metrics used to compare the different databases for their impact on taxonomy generation reflect the above issues. These metrics focus on counting the differences in unique phrases contained in each database. The metrics do not address the significance of the absent phrases, since that requires judgement about the quality of the application. The significance of the phrase absences from any database relative to other databases is judged qualitatively.

2) Higher Level Comparison

The high technical content phrases in each database are aggregated into clusters, and the clusters are integrated to form a taxonomy. The taxonomies are then compared for structural and conceptual differences. Two statistical approaches for taxonomy generation are used, factor matrix and multi-link clustering.

Since each phrase, phrase cluster, and taxonomy addresses some aspect of Raynaud's Phenomenon, an overview of Raynaud's Phenomenon will be presented before discussing the results. Because the main Raynaud's terminology used in the literature is not consistent (in many cases, Raynaud's Phenomenon is used interchangeably with Raynaud's Disease or Raynaud's Syndrome), the overview will include the distinction among these Raynaud variants.

Raynaud’s Phenomenon Overview

Raynaud’s Phenomenon is a condition in which small muscular arteries and arterioles, most commonly in the fingers and toes, go into spasm (contract) and cause the skin to turn pale (blanching) or a patchy red (rubor) to blue (cyanosis). While this sequence is normally precipitated by exposure to cold, and resolves with subsequent re-warming, it can also be induced by anxiety or stress. Blanching represents the ischemic (lack of adequate blood flow) phase, caused by digital artery vasospasm. Cyanosis results from de-oxygenated blood in capillaries and venules (small veins). Upon re-warming, a hyperemic phase ensues, causing the digits to appear red.

Raynaud’s Phenomenon can be a primary or secondary disorder. When Raynaud’s symptoms appear alone without an apparent underlying medical condition, it is referred to as Primary Raynaud’s Phenomenon or, formerly, as Raynaud’s Disease. In this condition, the blood vessels appear anatomically normal after the ischemic events. When an identifiable cause or a specific associated disease accompanies Raynaud’s symptoms, it is referred to as Secondary Raynaud’s Phenomenon. The auto-immune disorders, or conditions in which a person produces an immune response against his or her own tissues, are the typical medical conditions associated with Secondary Raynaud’s. In these cases, Raynaud’s Phenomenon may be more serious than in Primary Raynaud’s and may result in blood vessel scarring and long-term consequences. When Raynaud’s Phenomenon is associated with occupational activities, such as vibrating machinery or repetitive activity, it is often referred to as Occupational Raynaud’s. Similarly, Secondary Raynaud’s may be precipitated by exposure to harmful chemical compounds such as vinyl chloride, or to toxic therapeutic agents such as certain cancer chemotherapy drugs.

Thus, while the symptoms and signs of Raynaud’s Phenomenon occur as a direct consequence of reduced blood flow due to reversible blood vessel constriction, the underlying etiology may be a function of several parameters that affect blood flow. These include:

*Inflammation from the auto-immune disorders that can cause swelling and thereby constrict blood vessels;
*Increased sympathetic nervous system activity, that can affect the timing and duration of the blood vessel muscular contractions that cause constriction;
*Heightened digital vascular reactivity to vaso-constrictive stimuli, that causes the blood vessels to over-react and over-contract;
*Deposits along the blood vessel walls that can reduce blood flow and increase the flow sensitivity to contraction stimuli;
*Blood rheological properties that offer additional resistance to blood flow, and magnify the impact of blood vessel constriction;
*Blood constituents and hormones that can act as vaso-constrictors or vaso-dilators.

RESULTS

Phrase frequencies were generated for ten years of Mesh Terms, Titles, Mesh-Titles, and Abstracts from Medline records focused on Raynaud's Phenomenon (1975-1985). A sampling across frequency bands in the larger text fields showed that about 2/3 of the phrases could be classified as high technical content.

1) Quantitative and Qualitative Properties of Phrases in each Field

Table 1A lists the number of phrases in each field, incorporating single, double, and triple word phrases. The first column represents the databases in which the phrases appeared. The Abstracts contain almost an order of magnitude more phrases than the other two fields. Also, the low number of separate Mesh phrases reflects the restrictions imposed by a controlled vocabulary: less diversity, more uniformity. About 90% of the Abstract phrases tend to be unique, whereas slightly less than half of the Title and Mesh phrases are unique.

TABLE 1A – NUMBER OF PHRASES IN EACH FIELD

FIELD	TOTAL	UNIQUE

Abstract	44,029	40114
Title	5941	2780
Mesh	2735	1237
Mesh + Title	7950	

TABLE 1B – 'UNIQUE' TITLE PHRASES AND ABSTRACT REPRESENTATIONS

TITLE	ABSTRACT
ANGIOLOGIC	ANGIOLOGICAL
HAEMORHEOLOGY	HAEMORHEOLOGICAL
FINGER ULCER	FINGER ULCERATIONS
DISEASE GROUPS	DISEASE GROUP
ARTERIAL INFUSION	INTRA-ARTERIAL (IA) INFUSION
ANTIBODIES AGAINST SCL 70	ANTIBODIES AGAINST SCL-70
PERIPHERAL CIRCULATORY DISORDERS	PERIPHERAL CIRCULATION DISORDER

At this point, some readers may question the entries in Table 1A that show unique Title phrases relative to the Abstracts. Wouldn't every important technical word/ phrase in the Title appear in the Abstract? The answer depends on the definition of unique. If 'unique' is defined as 'identical', the answer is no. If 'unique' is defined as 'very similar', the answer is yes. For example, Table 1B contains some of the 'unique' Title phrases (as defined in Table 1A), and their Abstract representations.

Table 2 lists the 20 highest frequency high technical content phrases for each of the four databases (fields). All four databases share the following phrases: Raynaud, Disease, Systemic, Blood, Scleroderma, Antibodies, Finger(s), Skin, Lupus Erythematosus, Connective Tissue, and Vibration. The phrases are mainly single word, with some double word included. The high frequency phrases are relatively simple and generic, and major differences among the fields are not evident at this high frequency level of description.

TABLE 2 – HIGHEST FREQUENCY NON-DUPLICATIVE HIGH TECHNICAL CONTENT PHRASES

ABSTRACT PHRASES	MESH PHRASES	TITLE PHRASES	TITLE + MESH PHRASES
PATIENTS	HUMAN	RAYNAUD	RAYNAUD
RAYNAUD	DISEASE	DISEASE	DISEASE

DISEASE	RAYNAUD	PHENOMENON	HUMAN
PHENOMENON	BLOOD	SYNDROME	BLOOD
SYNDROME	COMPLICATIONS	SYSTEMIC	COMPLICATIONS
SYSTEMIC	THERAPY	PATIENTS	DRUG
BLOOD	DIAGNOSIS	SCLERODERMA	THERAPY
SCLERODERMA	IMMUNOLOGY	TREATMENT	DIAGNOSIS
TREATMENT	ADULT	SCLEROSIS	IMMUNOLOGY
ANTIBODIES	ETIOLOGY	TISSUE	ETIOLOGY
SKIN	MIDDLE	CONNECTIVE	DISEASES
TISSUE	DISEASES	LUPUS	MIDDLE
COLD	PATHOLOGY	BLOOD	SYSTEMIC
SYMPTOMS	PHYSIOPATHOLOGY	ERYTHEMATOSUS	PATHOLOGY
VASCULAR	THERAPEUTIC	ANTIBODIES	THERAPEUTIC
FLOW	SYSTEMIC	VIBRATION	SCLERODERMA
FINGER	SUPPLY	VASCULAR	SUPPLY
SCLEROSIS	SCLERODERMA	PERIPHERAL	SYNDROME
CONNECTIVE	SUPPORT	PRIMARY	SUPPORT
TEMPERATURE	RADIOGRAPHY	DISEASES	ADVERSE

Table 3 lists single word phrases in each database that are unique; i.e., not contained in any of the other databases. The Title-Abstract caveats with respect to uniqueness apply here as well.

TABLE 3 – UNIQUE PHRASES IN EACH DATABASE

	ABSTRACT	**TITLES**	**MESH**
1	ACETYLSALICYLIC	ADENINE	ABORTION
2	ATHEROSCLEROSIS	CRYOABLATION	BIOMECHANICS
3	ALANINE	EFAMOL	CORTEX
4	ANOREXIA	MALEATE	CRYOSURGERY
5	ANTIGLOBULIN	PREDILECTION	DYSPEPSIA
6	APERISTALSIS	WEARDALE	EMBRYOLOGY
7	CORTISOLAEMIA	SULPIRIDE	HYPOTHERMIA
8	ENDANGIITIS	NIACIN	ACETYLGLUCOSAMINIDASE
9	FLUORESCEIN-CONJUGATED	KAPOSI	ADENOSINE
10	HEXOPAL	FUROSEMIDE	ADULTORUM

The Mesh phrases, on average, appear somewhat more generic than the Title or Abstract phrases. They represent a controlled vocabulary, and can contain additional information to that present in Title or Abstract fields (and vice versa).

Table 4 contains phrases shared by all fields. These are the standard well-known phrases associated with Raynaud's, especially the more focused multi-word phrases.

TABLE 4 – SHARED PHRASES AMONG ALL FIELDS

	SINGLE WORDS	DOUBLE WORDS
1	ABDOMINAL	ACANTHOSIS NIGRICANS
2	TOXOPLASMOSIS	CONNECTIVE TISSUE
3	ACANTHOSIS	FACIAL HEMIATROPHY
4	CORPUSCLES	MULTIPLE SCLEROSIS
5	FIBROSIS	PLASMA EXCHANGE
6	GUANETHIDINE	RESPIRATORY FUNCTION
7	LEUKOCYTES	SCLERODERMA SYSTEMIC
8	METOPROLOL	SKIN MANIFESTATIONS
9	ORTHOSTATIC	THROMBOANGIITIS OBLITERANS
10	SUBCUTANEOUS	VASCULAR DISEASES

2) Taxonomies for all Four Databases

The previous section has shown differences among the databases at the micro, or individual phrase, level. Both the differences in total number of phrases, and number of unique phrases, were shown. For a crude measure of significance of these phrase differences, some examples of unique phrases were also shown.

However, differences at the macro, or aggregated phrase, level were not shown. These aggregated phrases, or clusters, can be thought of as concepts. It would be useful to ascertain whether there are conceptual differences among the databases.

There are two measures of importance in measuring conceptual differences. One measure is difference in structure of the overall database taxonomy, i.e., are there any concepts missing from any of the databases, and even if not, do all the concepts bear the same relationships across databases? The other measure is differences in resolution of each concept, i.e., how does the information content in each cluster vary across the different databases?

This section addresses the taxonomy structure metric. Two approaches are used to develop taxonomies for each database. A factor matrix is used first, followed by a multi-link clustering method. The combination has been used in previous text mining studies (e.g., 20), although the present application uses a synergistic combination of the two methods that offers substantial improvement in the quality of the resultant clusters. The factor matrix, which imports a relatively large number of words, is used as a filter for the words subsequently input to the clustering algorithms. This results in a selection of context-dependent words for input to the clustering algorithm, and produces relatively well defined clusters compared to the context-independent methods. In addition, the present application uses single words for clustering, rather than the multi-word phrases of previous applications. While some of the technical detail is lost by excluding the ordering information contained in multi-word phrases, inclusion of all single words compensates for the elimination of multi-word phrases due to the selection algorithm of the Natural Language Processor.

2A. Factor Matrix Filtering

Factor analysis aims to reduce the number of variables in a system, and detect structure in the relationships among variables. One of the key challenges in factor analysis is defining the number of factors to select. Different approaches have been suggested in the literature, but the two most widely used are the Kaiser criterion (21-22), and the Scree test (23). The Kaiser criterion states that only factors with eigenvalues greater than unity should be retained, essentially requiring that a factor extracts at least as much variance as the equivalent of one original variable. The Scree test plots factor eigenvalue (variance) vs factor number, and recommends that only those factors that extract substantive variance be retained. Operationally, the factor selection termination point becomes the ‘elbow’ of the plot, the point where the slope changes from large to small.

In most previous studies performed by the first author, the Kaiser criterion has been used to select the number of factors for the factor matrix. These previous studies have used an Excel add-in to generate the factor matrices and, due to Excel’s limitations on columns, have been limited approximately to 250 x 250 correlation matrices, or 250 words. The Kaiser criterion has yielded factor numbers in the range of 20-45, considered a reasonable number for analysis. However, in the present validation study, another software package that did not require Excel was used (TechOasis), and many more words were used for the correlation matrix. The Kaiser criterion yielded hundreds of factors, a number far too large for detailed factor analysis, and of questionable utility, since many of the eigenvalues were not too different from unity. The Scree Plot was examined, and used to select the number of factors for analysis.

Once the desired value of the Scree Plot 'elbow' has been determined, and the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (e.g., a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context.

In the factor matrix used, the rows are the words and the columns are the factors. The matrix elements Mij are the factor loadings, or the contribution of word i to the theme of factor j. The theme is determined by those words that have the largest absolute values of factor loading. Each factor had a positive value tail and negative value tail. For each factor, most of the time, one of the tails dominated in terms of absolute value magnitude. This dominant tail was used to determine the central theme of each factor.

For the first step in the factor matrix filtering process, the factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the dominant tail of each factor. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (24), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). A few words were eliminated manually, based on factor loading and estimate of technical content. Basically, any word that did not have a high absolute value of factor loading for at least one factor was eliminated from the subsequent clustering.

Before the clustering is described, the factors that formed the basis of the factor matrix filtering for all four fields will be described. Table 5 lists the number of words that were input to the factor matrix algorithm for each field. In each case, addition of succeeding lower frequency words (from the raw data word list ordered by frequency) made the factor matrix difficult to generate. A detailed description of all the factor and cluster results is presented in Appendix 2.

TABLE 5 – NUMBER OF WORDS INPUT TO FACTOR MATRIX

ABSTRACT	TITLE	MESH	M+T
659	428	465	519

Table 6 lists the different factor themes that can be extracted from all four fields, and identifies the factor(s) in which the theme occurs for each field. A more expanded description of the non-redundant themes is presented in Appendix 1.

TABLE 6 – FACTOR THEMES ASSOCIATED WITH EACH RECORD FIELD

FACTOR THEME	RECORD FIELD			
	ABST	TITLE	MESH	M+T
ANTIBODIES AND AUTOIMMUNE DISEASES	1	6	1	1
RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES	14		8	6
SYSTEMIC LUPUS ERYTHEMATOSUS CLASSIFICATION		5		
ADRENAL CORTEX HORMONES FOR LUPUS			5	5
SCLEROTIC AUTOIMMUNE DISEASES	6		13	
ENDOTHELIAL CELL ACTIVITY IN PSS		10		
MIXED CONNECTIVE TISSUE DISEASE		9		
CIRCULATING IMMUNE COMPLEXES	12	8	3	11
DEFICIENCIES OF COMPLEMENT COMPONENTS		13		
FASCIA-FOCUSED INFLAMMATION	13		1	13
CORONARY CIRCULATION AND HYPERTENSION	4		10	
SMOOTH MUSCLE RESPONSE TO DRUGS			14	14
PLASMA LIPID FRACTION CONTROL			11	3
CARDIAC BLOOD SUPPLY OBSTRUCTIONS		11		
DOUBLE-BLIND VASODILATOR TRIALS	2	4	2	6, 9
CALCIUM CHANNEL BLOCKERS			9	12
BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION	10		6	10
REDUCED PLATELET AGGREGATION VASODILATORS	5	12	2, 8	2, 8
PERIPHERAL CIRCULATORY SYSTEM VASODILATION	11			

FINGER BLOOD FLOW MEASUREMENTS	8	3		
NAIL-FOLD CAPILLARY MICROSCOPY	9	2		11
TRANSCUTANEOUS NERVE STIMULATION/VASODILATION		1		14
SURGICAL/ NERVE BLOCK CIRCULATION OBSTRUCTION SOLUTIONS	7		7, 12	7
CARPAL TUNNEL SYNDROME		14		
VIBRATION-BASED CIRCULATION PROBLEMS	3	7	13	9, 13
VINYL CHLORIDE EFFECTS ON BLOOD FLOW AND PROPERTIES	14	7	7, 12	4
TREATMENTS FOR CHEMICALLY-INDUCED NEOPLASMS			4	

At the highest level, the themes can be divided into two categories, auto-immunity and circulation. This division into these two categories is suggested by the data in the above table, is shown more sharply by the multi-link hierarchical structures, and reflects the experience of medical practice for Raynaud's Phenomenon. The first ten themes fit into the auto-immunity category, and the remainder fit into the circulation category.

At the next hierarchical level, the auto-immunity category can be divided into antibody and inflammation sub-categories, but other divisions are possible as well. The circulation category can be sub-divided into coronary vessel circulation and peripheral vessel circulation.

All fields are represented in some themes of each of the four second-level categories, although all fields are not represented in all themes. This should not be interpreted that a field-theme combination that is absent from the matrix means that the theme is absent from the field. It only means that, within the resolution afforded by a fourteen factor matrix, a specific theme was not among the fourteen major themes for that field.

For example, FASCIA-FOCUSED INFLAMMATION is not listed as a Title theme on the matrix. However, EOSINOPHILIC FASCIITIS appears in the Title words used to form the factor matrix. Also, BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION is not listed as a Title theme in the matrix, but BIOFEEDBACK appears in the Title word list. To fill in the blanks on Table 6, matrices with more factors would have to be generated, in turn producing a larger version of Table 6 with additional blanks.

Because of the smaller number of Title words, the themes in each category in which the Title is represented tend to be more specific relative to those of the Mesh or Abstract. The larger numbers of words in the Abstract or Mesh fields allow more generalized themes to populate the categories.

2B. Multi-Link Hierarchical Clustering

The filtered and conflated words resulting from the factor matrix filtering were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, from the resulting dendrogram (a tree-like structure that shows how the individual words cluster into groups in a hierarchical structure) were analyzed, and compared for the four fields. In most cases, only the top three hierarchical levels of each field's taxonomy were described.

First Taxonomy Level

The Abstract, Mesh, and Mesh + Title fields have the same first level taxonomy categories, auto-immunity and circulation, and these two categories are sharply delineated for all three fields.

The hierarchical structure of the Titles dendrogram is very different from that of the Abstracts dendrogram. When the Abstracts clusters divide at each lower level in the hierarchy, they split into sub-clusters that are of reasonably similar magnitude (usually, not always) and usually have complementary themes. When the Titles clusters divide, they are splitting into sub-clusters of very different magnitudes with themes that are not very complementary.

For example, the top level split of the Abstracts dendrogram is into two clusters containing 90 and 162 words. The themes of these clusters are auto-immunity and circulation, respectively, the two main complementary themes of Raynaud's Phenomenon. The top level split of the Titles dendrogram is into two clusters containing 4 and 249 words. The themes of these clusters are hereditary deficiency for the smaller cluster, and auto-immunity and circulation for the larger cluster. The next level split for the circulation cluster in the Abstracts dendrogram is into two clusters containing 25 and 137 words. The themes of these clusters are coronary circulation and peripheral circulation, respectively, the two main complementary themes of circulation. The next level split for the auto-immunity and circulation cluster in the Titles dendrogram is into two clusters containing 25 and 224 words. The themes of these clusters are controlled double blind trials of coronary vasodilators and antihypertensives, and autoimmunity and peripheral circulation and other aspects of coronary circulation, respectively. Thus, rather than splitting high level clusters into the next level clusters in a hierarchically conceptual process, the Titles dendrogram is stripping out a low level very detailed cluster from a high level cluster division. This type of structure has been found in studies of applications of a particular research or technology discipline, where the diversity of the applications produces numerous relatively unrelated themes. For the Titles, the relatively low word frequencies produce a sparse co-occurrence matrix, and the resultant clusters appear fragmented. The highest level categorizations in the Title field are not viewed as meaningful, and will not be addressed until the third level clusters are discussed.

Second Taxonomy Level

Table 7 shows all the second level categories from the Abstract, Mesh, and Mesh + Title fields, and the representation of each field in each category. Three asterisks denote fully applicable, while two asterisks denote a partial match.

TABLE 7 – SECOND LEVEL CATEGORY THEMES IN EACH FIELD

CATEGORY	ABST	MESH	M+T
AUTOIMMUNE DISEASES/ ANTIBODIES	***	***	***
INFLAMMATION/ FASCIA	***	***	
PERIPHERAL VESSEL CIRCULATION	***	**	
CORONARY VESSEL CIRCULATION	***	**	
PERIPHERAL AND CORONARY VESSEL CIRCULATION	**	***	
CHEMICALLY/ CHEMOTHERAPEUTICALLY-INDUCED DISEASES		***	***
NON-PSYCHOLOGICAL CIRCULATION TREATMENTS			***
PSYCHOLOGICAL CIRCULATION TREATMENTS			***

In all three fields, ‘auto-immune diseases and antibodies’ is the only common category. Interestingly, it was the first factor in the three factor matrices as well. The main difference between the Mesh and Abstract fields at this second hierarchical level is that the Mesh gives more recognition to consequences of vinyl chloride poisoning from industrial contact, and consequences of anti-neoplastic agents. Addition of the Title field to the Mesh field has the further effect of providing more recognition to psychologically-based treatments for improving circulation (biofeedback and autogenic training).

Lowest Level-Elemental Clusters

Table 8 shows all the third level categories and sub-themes from the Abstract, Title, Mesh, and Mesh + Title fields, and the representation of each field in each category. Three asterisks denote fully applicable, while two asterisks denote a partial match.

TABLE 8 – THIRD LEVEL CATEGORY THEMES IN EACH FIELD

THIRD LEVEL CLUSTER THEMES AND SUB-THEMES

CATEGORY	ABST	TITLE	MESH	M+T
ANTIBODIES	***	***	***	***
SCLEROTIC AUTOIMMUNE DISEASES	***	***	***	***
RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES	***	**	***	***
CREST SYNDROME AUTOIMMUNE DISEASES	***	***	***	***
CIRCULATING IMMUNE COMPLEXES	***	**	***	***
LIVER ABNORMALITIES IN RHEUMATIC DISEASES	**	***	***	***
FASCIA-RELATED INFLAMMATION	***	***	***	***
DOUBLE-BLIND CLINICAL TRIALS FOR VASODILATORS	***	***	***	***
VASODILATORS FOR REDUCED PLATELET AGGREGATION	***	***	***	***
TRANSCUTANEOUS NERVE STIMULATION		***		***
ARTERIAL OCCLUSIONS IN EXTREMITIES	**	***	***	***
FINGER BLOOD FLOW AND TEMPERATURE MEASUREMENTS	***	**	***	***
VIBRATION EFFECTS ON NERVOUS SYSTEM AND CIRCULATION	***	***	***	***
VINYL CHLORIDE EFFECTS ON NERVOUS SYSTEM AND CIRCULATION	***	***	***	***
CHEMOTHERAPY TOXICITY	**	***	***	***
NAILFOLD CAPILLARY MICROSCOPY	***	***		
CARDIOVASCULAR/ PULMONARY CIRCULATION PROBLEMS	***	**	***	***
BIOFEEDBACK AND AUTOGENIC TRAINING	***	*	***	***

Most of the sub-themes are covered in all fields, although there are some notable absences, and they tend to occur in the circulation category. TRANSCUTANEOUS NERVE STIMULATION is identified in Title and Mesh + Title, but not in Abstract or Mesh, while NAILFOLD CAPILLARY MICROSCOPY is specifically identified in Abstract and Title, but not in Mesh or Mesh + Title.

DISCUSSION AND CONCLUSIONS

Four fields (Abstract, Title, Mesh, Mesh + Title) in 932 Medline records were compared for information content. Four metrics were used to assess information content: 1) total number of phrases; 2) number of unique phrases; 3) factors; 4) clusters.

The Abstract field contains almost an order of magnitude more phrases than the other fields, with the difference becoming more exascerbated as the words per phrase increases. Even though the Mesh field tends to be larger than the Title field in volume, it has about half the total number of phrases. This is due to the controlled vocabulary limiting the phrase diversity relative to the unrestricted vocabulary of the Titles.

The difference in number of unique phrases between the Abstract field and the Title or Mesh fields is even more pronounced. These large differences are not evident when the high frequency phrases in each field are compared. At this end of the spectrum, many of the phrases tend to be the more generic representations of Raynaud's Phenomenon, and many are shared in common. The phrase differences tend to be more evident and pronounced at the lower frequency end of the spectrum.

In the theme comparison phase of the study, fourteen factors were used based on Scree plots and standardization. The Abstract field allowed the largest number of words to generate the factor matrices, while the Title field allowed the smallest number of words. These numbers differed by about 50%.

About 27 separate, but not unique, themes could be discerned from all the factors combined. These themes could be placed in two major categories, with two sub-categories per major category: Auto-immunity (antibodies, inflammation) and circulation (peripheral vessel circulation, coronary vessel circulation). All four sub-categories included representation from each field. Thus, while the focus of the representation of each field in each sub-category was moderately different, the four sub-category structure could be identified by analyzing the total factors in each field.

In the cluster comparison phase of the study, the phrases used to create the clusters were the most important phrases identified for each factor. Thus, the factor matrix served as a filter for words used for clustering. All clusters were based on about 250 words. A hierarchical multi-link aggregation clustering technique was used to form the clusters.

While clusters were generated for all four fields, the Title hierarchy tended to be fragmented due to sparsity of the co-occurrence matrix that underlies the clusters. Therefore, the Title clusters were examined at only the lower levels of aggregation.

The Abstract, Mesh, and Mesh + Title fields had the same first level taxonomy categories, auto-immunity and circulation. At the second level, the Abstract, Mesh, and Mesh + Title fields had the auto-immune diseases and antibodies sub-category in common. The Abstract and Mesh fields shared fascia inflammation as the other auto-immunity sub-category, while the other Mesh + Title sub-category focuses on vinyl chloride poisoning from industrial contact, and consequences of antineoplastic agents.

This latter difference illuminates previous observations that the different record fields sometimes operate at different meta-levels of description. Many antineoplastic agents can produce specific organ inflammation, myocutaneous inflammation, or systemic inflammation during the course of treatment. This effect can be referenced as inflammation, antineoplastic agents, or both. While the theme can be different superficially in the different fields, as in the present case, underneath the theme/ concept may be the same, but expression occurred at different meta-levels.

At the next level of categorization, most of the sub-categories are covered in all fields. There are some notable absences, and they tend to occur in the circulation category. For example, TRANSCUTANEOUS NERVE STIMULATION is identified in Title and Mesh + Title, but not in Abstract or Mesh, while NAILFOLD CAPILLARY MICROSCOPY is specifically identified in Abstract and Title, but not in Mesh or Mesh + Title.

Finally, what is the importance of the differences among the fields summarized in the present study? All levels of text mining, ranging from standard information retrieval to the more exotic literature-based discovery, tend to access records through phrase matching. As the results imply, there could be substantial differences in numbers and types of records retrieved, depending on which fields are accessed by the search engines.

For high level taxonomy generation, the field differences are less severe, but when lower level taxonomic detail is required, then the differences become important. The Title and Mesh fields have limited numbers of different phrases at the lower frequencies relative to the Abstracts, and this translates in differences in diversity of detail possible. For literature-based discovery in particular, access to related and disparate literatures will be limited due to sheer phrase volume and diversity limitations. The predominant publishing group in literature-based discovery (4, 6) has used Title and Keyword phrases for information processing almost exclusively. Use of Abstracts should result in much more literature content accessed.

REFERENCES

1. Hearst, M. A. Untangling Text Data Mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26, 1999.
2. Trybula, W.J. Text Mining. Annual Review of Information Science and Technology. 34. 385-419. 1999
3. Losiewicz, P., Oard, D., and Kostoff, R. N. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 15. 99-119. 2000.
4. Swanson, D.R, and Smalheiser, N.R. Implicit Text Linkages Between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery. Library Trends. 48 (1): 48-59. 1999.
5. Kostoff, R. N. Science and Technology Innovation. Technovation. 19:10. 593-604. 1999.
6. Swanson, D.R. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine. 30 (1): 7-18. 1986.
7. Gordon, M.D., Lindsay, R.K. Toward Discovery Support Systems: A Replication, Re-Examination, and Extension of Swanson's Work

On Literature-Based Discovery of a Connection Between Raynaud's And Fish Oil. Journal of the American Society for Information Science. 47 (2): 116-128. 1996

8. Weeber, M, Klein, H, de Jong-van den Berg L..W., Vos, R. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. Journal of the American SocietyfFor Information Science and Technology. 52 (7): 548-557. 2001.
9. Hersh, W.R., Hickam, D.H. A Comparison of Retrieval Effectiveness for 3 Methods of Indexing Medical Literature. American Journal of the Medical Sciences. 303 (5): 292-300 May 1992.
10. Su, K.C, Ries, J.E, Peterson, G.M, Sievert, M.C, Patrick, T.B, Moxley, D.E, Ries, L.D. Comparing Frequency of Word Occurrences in Abstracts and Texts Using Two Stop Word Lists. Journal of the American Medical Informatics Association. 682-686 Suppl. S 2001.
11. Johnson, E.D., Sievert, M.C. and McKinin, E.J. Retrieving Research Studies: A Comparison of Bibliographic and Full-Text Versions of the *New England Journal of Medicine.* Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care, 1995, 846-850.
12. Srinivasan, P. MeSHmap: A Text Mining Tool for MEDLINE. Journal of the American Medical Informatics Association. 642-646 Suppl. S 2001.
13. Healey, P., Rothman, H., and Hoch, P. An Experiment in Science Mapping for Research Planning. Research Policy. Vol. 15. 1986.
14. Funk, M.E., and Reid, C.A. Indexing Consistency in MEDLINE. Bull Med Libr Assoc. 71 (2). 1983. 176–183.
15. Lancaster, F.W. *Vocabulary Control for Information Retrieval.* Information Resource Press, 1972.
16. Lindberg, D.A., Humphreys, B.L. Computer Systems that Understand Medical Meaning. In: Scherrer, J.R, Cote, R.A, Mandil, S.H, editors, Computerized Natural Medical Language Processing for Knowledge Representation. Proceedings of the IFIP-IMIA WG6 International Working Conference; 1988 Sep 12-15; Geneva, Switzerland. Amsterdam: North-Holland; 1989. p. 5-17.
17. Humphreys, B.L, Schuyler, P.L. The Unified Medical Language System: Moving Beyond the Vocabulary of Bibliographic Retrieval.

In: Broering NC. editor. High-Performance Medical Libraries: Advances in Information Management for the Virtual Era. Westport (CT): Meckler; 1993. p. 31-44.

18. Hersh, W.R, Hickam, D.H, Haynes, R.B, McKibbon, K.A. A Performance and Failure Analysis of Saphire with a Medline Test Collection. Journal of the American Medical Informatics Association. 1 (1): 51-60 Jan-Feb 1994.
19. Mijnhout, G.S, Hooft, L, van Tulder M.W, Deville, W.L.J.M, Teule, G.J.J, Hoekstra, O.S. How to Perform a Comprehensive Search for FDG-PET Literature. European Journal of Nuclear Medicine. 27 (1): 91-97 Jan 2000.
20. Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography. Journal of Power Sources. 110:1. 163-176. 2002.
21. Kaiser, H.F. The Application Of Electronic Computers To Factor Analysis. Educational and Psychological Measurement. 20: 141-151. 1960.
22. Jackson, J. E. A Users Guide to Principal Components. Wiley, New York. 569. 1991.
23. Cattell, R.B. The Scree Test for the Number of Factors. Multivariate Behavioral Research. 1. 245 –276. 1966.
24. Kostoff, R. N. The Practice and Malpractice of Stemming. JASIST. 54:10. June 2003.

.

APPENDIX 1 – SUMMARY FACTOR DESCRIPTIONS

ANTIBODIES AND AUTOIMMUNE DISEASES
Different types of autoantibodies, especially anti-nuclear, anti-centromere, and extractable nuclear, and their relation to auto-immune diseases such as MCTD and SLE.

RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES
Auto-immune diseases associated with Raynaud’s Phenomenon, such as CREST syndrome.

SYSTEMIC LUPUS ERYTHEMATOSUS CLASSIFICATION
Compares specificity of the American Rheumatism Association criteria for the classification of systemic lupus erythematosus.

ADRENAL CORTEX HORMONES FOR LUPUS
Role of glucocorticoids in treating lupus erythematosus.

SCLEROTIC AUTOIMMUNE DISEASES
Scerloderma-spectrum autoimmune diseases, especially the CREST syndrome, and especially in females.

ENDOTHELIAL CELL ACTIVITY IN PSS
Endothelial cell activity in patients with progressive systemic sclerosis.

MIXED CONNECTIVE TISSUE DISEASE
Fatal conditions associated with mixed connective tissue disease.

CIRCULATING IMMUNE COMPLEXES
Serum levels of circulating immune complexes (including cryoglobulins) and immunoglobulins, especially IgG and IgM, with some emphasis on their relation to biliary cirrhosis.

DEFICIENCIES OF COMPLEMENT COMPONENTS
Hereditary deficiencies of complement components

FASCIA-FOCUSED INFLAMMATION
Inflammation, especially of the fascia (eosinophilic fasciitis), and the steroids used to control the inflammation.

CORONARY CIRCULATION AND HYPERTENSION
Coronary circulation and blood pressure problems; action of adrenergic antagonists, such as propranolol, on beta-2 adrenergic receptors.

SMOOTH MUSCLE RESPONSE TO DRUGS
In-vivo and in-vitro animal responses of smooth muscle to pharmacological agents, especially norepinephrine.

PLASMA LIPID FRACTION CONTROL
Prevention of myocardial infarction, angina pectoris, and cardiovascular arrhythmia through control of plasma lipid fractions.

CARDIAC BLOOD SUPPLY OBSTRUCTIONS

Cardiac and cerebral blood supply obstructions in young women, including migraine.

DOUBLE-BLIND VASODILATOR TRIALS
Double-blind trials for vasodilators such as nifedipine, or of imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation.

CALCIUM CHANNEL BLOCKERS
Calcium channel blockers and vasodilators, including their effect on smooth muscle contraction and treatment of non-cardiac disorders such as asthma.

BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION
Use of biofeedback and autogenic training techniques to induce relaxation, reduce stress headaches, and raise temperatures through improved circulation.

REDUCED PLATELET AGGREGATION VASODILATORS
Administration of vasodilators to improve circulation. Focuses on double-blind clinical trials of imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation. Also, therapeutic administration of drugs, mainly vasodilators such as prostaglandin-E, to increase regional blood supply.

PERIPHERAL CIRCULATORY SYSTEM VASODILATION
Vasodilation of the peripheral circulatory system after immersion, and the role of calcium in this process.

FINGER BLOOD FLOW MEASUREMENTS
Blood flow, and associated finger systolic blood pressure and temperature measurements.

NAIL-FOLD CAPILLARY MICROSCOPY
Diagnostic uses of electron microscopy and nail-fold capillary microscopy.

TRANSCUTANEOUS NERVE STIMULATION/ VASODILATION
Mediators of skin vasodilation induced by transcutaneous nerve stimulation.

SURGICAL/ NERVE BLOCK CIRCULATION OBSTRUCTION SOLUTIONS
Surgical and nerve block solutions to remove motor system constrictions on circulation, including surgical treatments for subclavian artery lesions and embolism at the thoracic outlet and other arteries.

CARPAL TUNNEL SYNDROME
Carpal tunnel syndrome.

VIBRATION-BASED CIRCULATION PROBLEMS
Impact of vibratory tools (such as chain saws) on circulation, and adverse effects of occupational vibration on the peripheral nervous system.

VINYL CHLORIDE EFFECTS ON BLOOD FLOW AND PROPERTIES
Vinyl chloride effects on blood flow and properties, especially chemically-induced diseases, such as osteolysis, resulting from the industrial use of vinyl chloride compounds.

TREATMENTS FOR CHEMICALLY-INDUCED NEOPLASMS
Chemotherapeutic treatments for neoplasms, especially testicular neoplasms.

APPENDIX 2 – DETAILED TAXONOMIES

2A-Abstract Taxonomies

For the Abstracts, the text was converted to single words and their frequencies. After trivial word filtering, the highest frequency words were selected for input to the factor matrix generator. Addition of succeeding lower frequency words made the factor matrix difficult to generate.

2A1-Factor Matrix Taxonomy – Single Words

A fourteen factor matrix resulted from Scree Plot analysis of the Abstract words. Following is a brief narrative of each factor theme.

Factor 1 - different types of autoantibodies, especially anti-nuclear and extractable nuclear, and their relation to auto-immune diseases such as MCTD and SLE.

Factor 2 - double-blind trials for vasodilators such as nifedipine.

Factor 3 - impact of vibratory tools (such as chain saws) on circulation.

Factor 4 - coronary circulation and blood pressure problems.

Factor 5 - administration of vasodilators to improve circulation.

Factor 6 - scerloderma-spectrum autoimmune diseases, especially the CREST syndrome.

Factor 7 - surgical and nerve block solutions to remove motor system constrictions on circulation.

Factor 8 - blood flow, and associated finger blood pressure and temperature measurements.

Factor 9 - diagnostic use of nail-fold capillary microscopy.

Factor 10 - use of biofeedback training to reduce stress headaches, and raise temperatures through improved circulation.

Factor 11 - vasodilation of the peripheral circulatory system after immersion, and the role of calcium in this process.

Factor 12 - serum levels of circulating immune complexes and immunoglobulins, especially IgG and IgM.

Factor 13 - inflammation, especially of the fascia, and on the steroids used to control the inflammation.

Factor 14 - two tails of almost similar magnitude. The dominant tail focuses on autoimmune diseases associated with Raynaud's

Phenomenon, while the inferior tail focuses on vinyl chloride effects on blood flow and properties.

The fourteen factor matrix was then used for word filtering and selection. In the present study, the words in the factor matrix had to be culled to the approximately 250 allowed by the Excel-based clustering package, WINSTAT. The 250 word limit is an artifact of Excel. Other software packages may allow more or less words to be used for clustering, but all approaches perform culling to reduce dimensionality. The filtering process presented here is applicable to any level of filtered words desired.

The factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the tail of each factor. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (24), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). A few words were eliminated manually, based on factor loading and estimate of technical content.

2A2-Multi-Link Taxonomy – Single Words

The filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, from the resulting dendrogram (a tree-like structure that shows how the individual words cluster into groups in a hierarchical structure) will now be described. Only the top three hierarchical levels will be described.

The top hierarchical level can be divided into two major clusters. Cluster 1 focuses on autoimmunity, and cluster 2 focuses on circulation. The second hierarchical level can be divided into four clusters, where cluster 1 is divided into clusters 1a and 1b, and cluster 2 is divided into clusters 2a and 2b. Cluster 1a focuses on auto-immune diseases and antibodies, while cluster 1b focuses on inflammation, especially of the fascia. Cluster 2a focuses on peripheral vascular circulation, while cluster 2b focuses on coronary vascular circulation.

Most of the clusters in the second hierarchical level can be rationally divided into two sub-clusters, to produce the third hierarchical level clusters. Cluster 1a1 has multiple themes: different types of antibodies, especially anti-nuclear and extractable nuclear, and their relation to autoimmune diseases; sclerotic types of autoimmune diseases; autoimmune diseases associated with Raynaud's Phenomenon; and CREST syndrome autoimmune diseases. It incorporates the themes of factors 1, 6, and 14. Cluster 1a2 focuses on circulating immune complexes, and parallels the theme of factor 12. Cluster 1b is too small to subdivide further, and stops at the second hierarchical level. It parallels the theme of factor 13.

Cluster 2a1 has multiple themes: double-blind clinical trials for vasodilators; administration of vasodilators to reduce platelet aggregation and improve circulation; blood flow, and associated finger blood pressure and temperature measurements; and occupational exposures, mainly vibrating tools and vinyl chloride, that impact the peripheral and central nervous systems and subsequently impact circulation. It incorporates the themes of factors 2, 3, 5, 7, 8. Cluster 2a2 focuses on nailfold capillary microscopy as a diagnostic for for microcirculation, and parallels the theme of factor 9. Cluster 2b1 focuses on cardiovascular system problems, and parallels the theme of factor 4. Cluster 2b2 focuses on biofeedback training to reduce stress and headaches, and increase relaxation, and parallels the theme of factor 10.

2B-Title Taxonomies - Single Words

For the Titles, the text was converted to single words and their frequencies. After trivial word filtering, the highest frequency words were selected for input to the factor matrix generator. Addition of succeeding lower frequency words made the factor matrix difficult to generate. Because of the substantially smaller number of words in Titles relative to Abstracts, the frequencies are also much lower, and the cutoff frequency for factor matrix generation, and associated number of words used, is also much smaller.

2B1–Factor Matrix Taxonomy - Single Words

A seven factor matrix resulted from Scree Plot analysis of the Abstract words, but a fourteen factor matrix was selected for comparison purposes. The factor eigenvalue at cutoff was slightly above three, so there is still substantial variance being extracted by the marginal fourteenth factor.

Factor 1 - mediators of skin vasodilation induced by transcutaneous nerve stimulation.

Factor 2 - diagnostic uses of electron microscopy and nail-fold capillary microscopy.

Factor 3 - finger systolic blood pressure and temperature measurements.

Factor 4 - double-blind trials for thromboxane synthetase inhibitor vasodilators.

Factor 5 - compares specificity of the American Rheumatism Association criteria for the classification of systemic lupus erythematosus.

Factor 6 - different types of autoantibodies, especially antinuclear and extractable nuclear, and their relation to autoimmune diseases.

Factor 7 - workers exposure to vibrating tools and vinyl chloride.

Factor 8 - circulating immune complexes in primary biliary cirrhosis.

Factor 9 - fatal conditions associated with mixed connective tissue disease.

Factor 10 - endothelial cell activity in patients with progressive systemic sclerosis.

Factor 11 - cardiac and cerebral blood supply obstructions in young women, including migraine.

Factor 12 - two tails of almost similar magnitude, with an overall focus on pharmacology. The dominant tail focuses on use of prazosin to treat hypertension and congestive heart failure, while the inferior tail focuses on use of thromboxane synthetase inhibitor dazoxiben for platelet aggregation reduction.

Factor 13 - hereditary deficiencies of complement components.

Factor 14 - two tails of almost similar magnitude. The dominant tail focuses on correction of hemorheological disorders in patients with arterial occlusions, while the inferior tail focuses on carpal tunnel syndrome.

2B2-Multi-Link Taxonomy - Single Words

The filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, resulting from the dendrogram will now be described. The hierarchical structure of the Titles dendrogram is very different from that of the Abstracts dendrogram. When the Abstracts clusters divide at each lower level in the hierarchy, they split into sub-clusters that are of reasonably similar magnitude (usually, not always) and usually have complementary themes. When the Titles clusters divide, they are splitting into sub-clusters of very different magnitudes with themes that are not very complementary.

For example, the top level split of the Abstracts dendrogram is into two clusters containing 90 and 162 words. The themes of these clusters are autoimmunity and circulation, respectively, the two main complementary themes of Raynaud's Phenomenon. The top level split of the Titles dendrogram is into two clusters containing 4 and 249 words. The themes of these clusters are hereditary deficiency for the smaller cluster, and autoimmunity and circulation for the larger cluster. The next level split for the circulation cluster in the Abstracts dendrogram is into two clusters containing 25 and 135 words. The themes of these clusters are coronary circulation and peripheral circulation, respectively, the two main complementary themes of circulation. The next level split for the autoimmunity and circulation cluster in the Titles dendrogram is into two clusters containing 25 and 224 words. The themes of these clusters are controlled double blind trials of coronary vasodilators and antihypertensives, and autoimmunity and peripheral circulation and other aspects of coronary circulation, respectively. Thus, rather than splitting high level clusters into the next level clusters in a hierarchically conceptual process, the Titles dendrogram is stripping out a low level very detailed cluster from a high level cluster division. This type of structure has been found in studies of applications of a particular research or technology discipline, where the diversity of the applications produces numerous relatively unrelated themes. For the Titles, the relatively low word frequencies produce a sparse co-occurrence matrix, and the resultant clusters appear fragmented.

There is little point in presenting the hierarchical structure for the Titles dendrogram. Instead, the lowest level clusters equivalent in number to the third level of the Abstracts dendrogram will be presented. Cluster 1 focuses on autoimmune diseases constituting the CREST syndrome. It incorporates the theme of one of the tails of factor 14. Cluster 2 focuses on liver abnormalities in rheumatic diseases and associated immune complexes and other immunological markers, although un-related terms like BIOFEEDBACK are present. It parallels the theme of factor 8. Cluster 3 focuses on sclerosis-based autoimmune diseases, although circulation-related terms like PROSTACYCLIN are present. It incorporates the themes of factors 5 and 11. Cluster 4 focuses on arterial occlusions in the extremities, although weakly-related terms like CARDIOVASCULAR are present. It incorporates the theme of one of the tails of factor 14. Cluster 5 focuses on cardiac and thyroid toxicity from chemotherapy. It incorporates the theme of factor 11. Cluster 6 focuses on peripheral circulation problems, especially those induced by vibratory tools. It incorporates the themes of factors 3 and 7. Cluster 7 focuses on occupational exposures, mainly vinyl chloride and vibrating tools (e.g., chain saws) used by forestry workers. It parallels the theme of factor 7. Cluster 8 focuses on mixed connective tissue disease and underlying laboratory antibody tests and nailfold capillary tests. It incorporates the themes of factors 2, 6, and 9. Cluster 9 focuses on inflammation, especially of the fascia. Cluster 10 has a dual focus of transcutaneous nerve stimulation for vasodilation, and pharmaceutical intervention to reduce hypertension and increase vasodilation. It incorporates the themes of factor 1 and one tail of factor 12. Cluster 11 focuses on controlled double-blind clinical trials for drugs to reduce platelet aggregation, reduce hypertension, and serve as vasodilators. It incorporates the themes of factor 4 and one tail of factor 12. Cluster 12 focuses on calcium channel blockers, and is closely allied with cluster 11. It parallels the theme of factor 13.

2C-Mesh Taxonomies

For the Mesh, the text was converted to single words and their frequencies. After trivial word filtering, the highest frequency words were selected for input to the factor matrix generator. Addition of succeeding lower frequency words made the factor matrix difficult to generate.

2C1-Factor Matrix Taxonomy - Single Words

An approximately fifteen factor matrix resulted from Scree Plot analysis of the Mesh words, but a fourteen factor matrix was selected for comparison purposes. The factor eigenvalue at cutoff was above three, so there is still substantial variance being extracted by the marginal fourteenth factor.

Factor 1 - two tails of important magnitude. The dominant tail focuses on antibodies, especially anti-nuclear and anti-centromere, while the inferior tail focuses on inflammation of the fascia, especially eosinophilic fasciitis.

Factor 2 - focuses on double-blind clinical trials of imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation.

Factor 3 - immunoglobulins, especially IgG and IgM, and circulating immune complexes (including cryoglobulins), and their relation to biliary cirrhosis.

Factor 4 - chemotherapeutic treatments for neoplasms, especially testicular neoplasms.

Factor 5 - role of glucocorticoids in treating lupus erythematosus.

Factor 6 - use of biofeedback and autogenic training techniques to induce relaxation, reduce stress headaches, and raise temperatures through improved circulation.

Factor 7 - arteriopathies of occupational diseases, mainly the use of vinyl chloride compounds in industry and associated chemically-induced poisoning, as well as subclavian artery lesions and embolism at the thoracic outlet and other arteries.

Factor 8 - has two tails of important magnitude. The dominant tail focuses on therapeutic administration of drugs, mainly vasodilators such as prostaglandin-E, to increase regional blood supply, while the inferior tail focuses on CREST syndrome.

Factor 9 - calcium channel blockers and vasodilators, including their effect on smooth muscle contraction and treatment of non-cardiac disorders such as asthma.

Factor 10 - action of adrenergic antagonists, such as propranolol, on beta-2 adrenergic receptors .

Factor 11 - prevention of myocardial infarction, angina pectoris, and cardiovascular arrhythmia through control of plasma lipid fractions.

Factor 12 - two tails of important magnitude. The dominant tail focuses on chemically-induced diseases, such as osteolysis, resulting from the industrial use of vinyl chloride compounds, while the inferior tail focuses on surgical treatments for subclavian artery lesions and embolism at the thoracic outlet and other arteries.

Factor 13 - two tails of important magnitude. The dominant tail focuses on adverse effects of occupational vibration on the peripheral nervous system, while the inferior tail focuses on systemic scleroderma in females.

Factor 14 - in-vivo and in-vitro animal responses of smooth muscle to pharmacological agents, especially norepinephrine.

2C2-Multi-Link Taxonomy - Single Words

The filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, resulting from the dendrogram will now be described. Only the top three hierarchical levels will be described.

The top hierarchical level can be divided into two major clusters. Cluster 1 focuses on circulation, and cluster 2 focuses on autoimmunity. The second hierarchical level can be divided into four clusters, where cluster 1 is divided in clusters 1a and 1b, and cluster 2 is divided into clusters 2a and 2b. Cluster 1a focuses on peripheral and coronary circulation, while cluster 1b focuses on chemically-induced diseases (from vinyl chloride compounds) associated with Raynaud's symptoms and chemotheraputic agents (mainly for testicular cancer) associated with Raynaud's symptoms. Cluster 2a focuses on autoimmune diseases and antibodies, while cluster 2b focuses on inflammation, especially of the fascia.

Most of the clusters in the second hierarchical level can be rationally divided into two sub-clusters, to produce the third hierarchical level clusters. Cluster 1a1 has multiple themes: finger blood flow and supply measurements; vasodilator and anti-hypertensive agents; double-blind clinical drug trials; imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation; biosynthesized prostaglandins for reduced platelet aggregation and vasodilation; prevention of myocardial infarction, angina pectoris, and cardiovascular arrhythmia through control of plasma lipid fractions; action of adrenergic antagonists, such as propranolol, on beta-2 adrenergic receptors to reduce cardiac output; calcium channel blockers and vasodilators, including their effect on smooth muscle contraction and treatment of non-cardiac disorders such as asthma; responses of smooth muscle to pharmacological agents, especially norepinephrine; use of biofeedback and autogenic training techniques to induce relaxation, reduce stress headaches, and raise temperatures through improved circulation. It incorporates the themes of factors 2, 6, 8, 9, 10, 11, and 14. Cluster 1a2 focuses on the peripheral nervous system, including occupational vibration impact, and obstructions to peripheral circulation. It parallels the themes of factors 12 and 13.

Cluster 1b1 focuses on chemically-induced poisoning from industrial vinyl chloride compounds, and parallels the themes of factors 7 and 12. Cluster 1b2 focuses on neoplasms, mainly testicular, and antineoplastic drugs.

Cluster 2a1 focuses on auto-immune diseases, antibodies (mainly antinuclear), and circulating immune complexes. It incorporates the themes of factors 1, 3, 5, 8, and 13. Cluster 2a2 focuses on associations with pulmonary fibrosis such as reduced total lung capacity and dermatomyositis. Cluster 2b1 focuses on renal and cardiac manifestations of SLE, including reduced adrenal cortex hormone output, and parallels the sub-theme of factor 5. Cluster 2b2 focuses on inflammation of the fascia, and parallels the sub-theme of factor 1.

2D-Title+Mesh Taxonomies

For the Title + Mesh, the text was converted to single words and their frequencies. After trivial word filtering, the highest frequency words were selected for input to the factor matrix generator. Addition of succeeding lower frequency words made the factor matrix difficult to generate.

2D1-Factor Matrix Taxonomy – Single Words

An approximately sixteen factor matrix resulted from Scree Plot analysis of the Abstract words, but a fourteen factor matrix was selected for comparison purposes. The factor eigenvalue at cutoff was well above three, so there is still substantial variance being extracted by the marginal fourteenth factor.

Factor 1 - antibodies, especially antinuclear and anti-centromere, and the immuno-fluorescent technique for labeling specific antibodies with fluorescent dyes.

Factor 2 - imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation.

Factor 3 - blocking of alpha-adrenergic receptors by drugs such as prazosin and quinazoline derivatives to obtain more favorable plasma lipid profiles, and reduce hypertension, congestive heart failure, arrhythmia, angina pectoris, and myocardial infarction.

Factor 4 - chemically-induced diseases, such as neoplasms and osteolysis, resulting from the exposure of industrial workers to the use of vinyl chloride compounds

Factor 5 - lupus erythematosus, including the role of glucocorticoids in treating lupus erythematosus, the role of lupus erythematosus in relation to other connective tissue disease components, and the correlation of elevated antinuclear antibodies (particularly ribonucleoproteins) with lupus erythematosus.

Factor 6 - two tails of important magnitude. The dominant tail focuses on the CREST syndrome and its relation to pulmonary fibrosis and diffusing capacity, while the inferior tail focuses on double-blind clinical trials of drug therapies to increase blood supply.

Factor 7 - surgical and nerve block techniques to reduce upper extremity arterial occlusions and ischemia, mainly subclavian artery lesions and embolisms at the thoracic outlet and brachial and carotid arteries.

Factor 8 - therapeutic usage of vasodilators that inhibit platelet aggregation.

Factor 9 - two tails of important magnitude. The more generic dominant tail focuses on double-blind drug therapy clinical trials, while the more specific inferior tail focuses on the nervous system impact of occupational diseases using vibratory tools, and the consequent impact on microcirculation in the extremities.

Factor 10 - use of biofeedback and autogenic training techniques to induce relaxation and reduce stress headaches.

Factor 11 - two tails of important magnitude. The dominant tail focuses on immunoglobulins, especially IgG and IgM, and circulating immune complexes (including cryoglobulins), and their relation to biliary cirrhosis, while the inferior tail focuses on the use of finger nailfold capillary microscopy in the diagnosis of Raynaud's Phenomenon.

Factor 12 - calcium channel blockers and vasodilators, including their effect on treatment of non-cardiac disorders such as asthma.

Factor 13 - two tails of important magnitude. The dominant tail focuses on inflammation of the fascia, especially eosinophilic fasciitis, while the inferior tail focuses on the Raynaud's signs of reduced regional blood flow and supply to the fingers, especially resulting from occupational use of vibratory equipment.

Factor 14 - animal responses of smooth muscle to pharmacological agents, especially norepinephrine, and to electric transcutaneous nerve stimulation.

2D2-Multi-Link Taxonomy - Single Words

The filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, resulting from the dendrogram will now be described. Only the top three hierarchical levels will be described.

The top hierarchical level can be divided into two major clusters. Cluster 1 focuses on circulation, and cluster 2 focuses on autoimmunity. The second hierarchical level can be divided into four clusters, where cluster 1 is divided into clusters 1a and 1b, and cluster 2 is divided into clusters 2a and 2b. Cluster 1a focuses on non-psychological aspects of treating and improving circulation, while cluster 1b focuses on biofeedback and autogenic training for treating and improving circulation. Cluster 2a focuses on autoimmune diseases and antibodies, while cluster 2b focuses on consequences of vinyl chloride poisoning from industrial contact, and anti-neoplastic agents.

Most of the clusters in the second hierarchical level can be rationally divided into two sub-clusters, to produce the third hierarchical level clusters. Cluster 1a1 focuses on peripheral vascular circulation, and incorporates the themes of factors 7, 9, 13, and 14. Cluster 1a2 focuses on coronary vascular circulation, and incorporates the themes of factors 2, 3, 8, and 12. Cluster 1b is too small and focused to subdivide further, and stops at the second hierarchical level. It parallels the theme of factor 10.

Cluster 2a1 focuses on autoimmune diseases and antibodies, and incorporates the themes of factors 1, 5, 6, 11, 13. Cluster 2a2 focuses on pulmonary fibrosis and associated decrease in lung diffusing capacity and respiratory volume, and parallels the dominant sub-theme of factor 6. Cluster 2b1 focuses on chemotherapeutic agents for neoplasms, and parallels a sub-theme of factor 4. Cluster 2b2 focuses on chemically-induced non-carcinogenic diseases resulting from exposure of industrial workers to the use of vinyl chloride compounds, and parallels the theme of factor 4.